Proteomics goes parallel

Collins, B. C., & Aebersold, R. (2018). Proteomics goes parallel. *Nature Biotechnology*, *36*(11). https://doi.org/10.1038/nbt.4288

Published in:
Nature Biotechnology

Document Version:
Peer reviewed version

Proteomics goes parallel

Massively parallel sequencing of peptides could signal a new era of high-throughput proteomics.

Ben C. Collins[1] and Ruedi Aebersold[1, 2]

[1]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich and [2]Faculty of Science, University of Zurich, Zurich, Switzerland

Proteomics has yet to attain the power of genomics and transcriptomics. The impressive performance of technologies for nucleic-acid sequencing rests on massively parallel measurements of short oligonucleotides, using fluorescence as a readout. In this issue, Swaminathan *et al.*[1] demonstrate that parallel fluorescence sequencing is also achievable for peptides. Their innovative method combines elements of classic protein chemistry with features of the optical systems used in nucleic-acid sequencing. Although further optimization is needed, the study fascinates with the prospect of a generally accessible, reliable, and truly universal proteomic technology.

Proteins are indispensible to living systems in their roles as chemical catalysts, structural components, and mediators of physiological processes. The ability to accurately identify and quantify proteins would greatly contribute to the understanding of biology. Today, proteomes are frequently predicted or inferred from transcriptomes. It is well

documented that the dependency between protein and mRNA levels is complex, and that predicting one from the other is imprecise and unreliable[2]. Why then are necessarily imprecise predictions from mRNA preferred over direct protein measurements in many instances? The answer lies in the state and accessibility of the respective measurement techniques: whereas essentially complete transcriptome analysis is readily available to biologists via core facility and commercial providers, proteome analysis is still most effectively performed by expert labs and cannot easily reach the throughput, robustness and reproducibility of transcriptome analysis.

The first generation of DNA sequencers, which produced groundbreaking genome maps, was based on sequential sequencing of isolated DNA segments—an intrinsically slow and expensive process even with automation. Widely accessible genomic analysis became possible only with the development of methods that sequence millions of nucleic acid segments in parallel[3], allowing complete genomic maps to be generated at high throughput and coverage and at low cost. These commercially well-supported techniques have transformed biomedical research and become a mainstay of experimental biology.

Although 'top down' proteomics approaches are emerging[4], proteins have traditionally been quantified and sequenced using 'bottom up' methods. As in genomics, these methods analyze constituent segments—in this case, peptides generated by enzymatic cleavage of proteins. In the 1950s, Pehr Edman invented a cyclic process of chemical reactions, known as Edman degradation[5], to determine the amino acid sequence of peptides. It consists of the coupling of phenyl isothiocyanate to accessible amino groups followed by release of the derivatized N-terminal amino acid from the peptide chain, generating a new N-terminus. The released amino acid is identified, and the process is

repeated to establish the peptide sequence. The Edman process is slow and requires large amounts of highly purified peptides. Yet, essentially all protein sequences known until the early 1990s were determined with this process.

In the 1990s, mass spectrometry (MS) became the method of choice for protein sequencing, leaving Edman degradation in the realm of science history. MS techniques to infer protein identity and quantity from measurements of the mass to charge ratio and fragmentation pattern of peptide segments have become highly sophisticated, powerful and versatile, and thus widely used[6]. Emulating the path of genomics, these techniques have progressed from manual sequencing of specific oligomers, to automated, sequential sequencing of peptides at high throughput, to parallel sequencing of multiple peptides by means of data-independent analyses[7,8], exemplified by SWATH-MS[9]. Although their throughput, accuracy and reproducibility are remarkable, the goal of routine, complete proteome quantification of large sample cohorts, akin to genomic analyses, has remained elusive.

It is conceivable that continued advances within the current framework of data-independent-acquisition MS will eventually achieve a performance on par with genomics. But it is also possible that a full account of the complexity and depth of proteomes will require disruptive new technologies. Although nanopore sequencing of proteins has shown promise[10], the peptide fluorosequencing method of Swaminathan *et al.*[1] appears to be the most advanced example of such a disruptive approach with a clear path to routine use. It is a marriage across the ages—between the largely forgotten Edman degradation chemistry

and the principles of massively parallel-in-space fluorescence imaging developed for next-generation DNA sequencing **(Fig. 1)**.

The first step of the new method is to generate an array of sequencing substrates by fluorescently labeling peptides at specific amino acid side chains and immobilizing them at their C-termini in the flow cell of a sequencing system. The immobilized peptides are then subjected to Edman degradation steps in parallel, and after each step the ensemble of immobilized substrates is imaged. In contrast to classic Edman degradation, in which the phenylthiohydantoin–amino acid conjugates eliminated at each step are identified, the stepwise degradation serves simply as a register to measure the decrease of fluorescence intensity caused by elimination of a labeled amino acid. The sequence of each immobilized substrate is inferred by relating the constraints derived from the observed fluorescence patterns to a protein sequence database using a sophisticated software tool developed for this purpose.

In this study the authors have taken the first steps towards feasibility of peptide fluorosequencing. Specifically, they (i) describe an imaging system compatible with the harsh conditions associated with the Edman degradation chemistry, (ii) demonstrate determination of the precise position of fluorescently labeled lysine or cysteine residues in model peptides, (iii) characterize sources of error and inefficiencies in the system, (iv) simulate the potential to identify proteins from more complex proteomes and provide a computational framework to infer peptide sequences from the observed fluorescent patterns, and (v) demonstrate the localization of a particular phosphorylated serine residue from a peptide containing multiple serines.

The peptide fluorosequencing method of Swaminathan *et al.*[1] is exciting because it highlights a clear path toward peptide, and conceivably protein, sequencing at very high throughput and reproducibility and potentially low cost. A substantial advantage of the system is that it capitalizes on a collection of well-characterized processes from other strategies (Edman chemistry, massively parallel DNA sequencing, and MS-based computational strategies for sequence database searching) that may speed maturation from proof-of-concept to a routinely applicable method. Furthermore, the data generated by the method should bear some resemblance to the data produced by its massively parallel antecedents in the world of genomics and transcriptomics. This could accelerate the adoption of peptide fluorosequencing by the broader biological community, in contrast to MS-based proteomics technologies, whose uptake has arguably been slowed by their technical and computational difficulty.

As Swaminathan *et al.*[1] note, several technical and conceptual challenges must be overcome before the method can reach its full potential. The issues are mainly rooted in the nature of Edman chemistry and the complexity of the human proteome, and include the following: (i) even at the yield per degradation step shown in the paper (91-97%), the length of achievable peptide sequences is limited; (ii) because the sequencing yield is sequence dependent, challenging sequences, such as proline-rich stretches, may obscure the sharpness of the fluorescent patterns; (iii) the number of functional groups accessible to fluorescent labeling is limited to the chemically reactive groups in peptides, predominantly amino, carboxyl and sulfhydryl groups, thus capping the information content of the fluorescence patterns; (iv) modified residues will generally not be recognized unless they are specifically fluorescently labeled, and a specific labeling chemistry is known for only a

small subset of modifications; (v) the large dynamic range of the human cellular proteome (~10^7), along with the high number of peptides generated per protein by enzymatic digestion (~10^2) and the large number of open reading frames expressed per cell (~10^4) constitute an enormous analytical challenge, even disregarding proteoform diversity. For peptide fluorosequencing, meeting these challenges requires a level of substrate multiplexing that has not yet been achieved. Although the system implemented by the authors is limited to the analysis of relatively simple sample mixtures, the path forward seems well laid out and is certainly one worth taking.

1. Swaminathan, J. *et al.* Highly parallelized single molecule sequencing and identification of proteins. *Nat Biotech*
2. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165,** 535–550 (2016).
3. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17,** 333–351 (2016).
4. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem (Palo Alto Calif)* **9,** 499–519 (2016).
5. Edman, P. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chem. Scand.* **4,** 283–293 (1950).
6. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355 (2016).

7. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Meth.* **1,** 39–45 (2004).
8. Purvine, S., Eppel, J. T., Yi, E. C. & Goodlett, D. R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **3,** 847–50 (2003).
9. Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11,** O111.016717 (2012).
10. Robertson, J. W. F. & Reiner, J. E. The Utility of Nanopore Technology for Protein and Peptide Sensing. *PROTEOMICS* 1800026 (2018). doi:10.1002/pmic.201800026

Figure 1

Peptide fluorosequencing as described by Swaminathan *et al.*[1]. Complex peptide mixtures, most likely derived from enzymatic or chemical cleavage of protein extracts, are labeled with different fluorophores for each amino acid residue (left). In this case, we depict a 2-color scheme where lysine and cysteine residues are labeled with distinct fluorophores. The labeled peptides are immobilized at their C-terminus using amide linkage to aminosilanes on a glass cover slip. The peptides are then subjected to iterative cycles of cleavage of the N-terminal amino acid residue by the Edman degradation and fluorescence imaging (center). The fluorescence intensity at each location (i.e. peptide) is tracked as a function of Edman cycles. The pattern of fluorescence intensity drops is interpreted to provide a partial sequence annotation for each peptide, which can be matched and scored against a protein sequence database to infer the most likely set of proteins present in the sample (right).

Fluorescently label specific amino acids on peptides and immobilize

Image

Cleave N-term amino acid (Edman)

Sequence inference and database comparison

Peptide #1

Peptide #3

Peptide #2

NH_2

K

C

x ~10^6 peptides

Fluorescence Imaging

Fluorescence

Edman cycles

Peptide #1

Peptide #2

Peptide #3

X-K-K-C-X-X-K-...

C-X-K-X-X-K-...

X-K-...

Protein Sequence DB